OMICS A Journal of Integrative Biology
Volume 9, Number 2, 2005

The Tropical Biominer Project: Mining Old Sources for New Drugs

FRANÇOIS ARTIGUENAVE,[2,4] ANDRÉ LINS,[1] WESLEY DIAS MACIEL,[1]
ANTONIO CELSO CALDEIRA JUNIOR,[1] CARLA NACIF-COELHO,[4]
MARIA MARGARIDA RIBEIRO DE SOUZA LINHARES,[4]
GUILHERME CORREA DE OLIVEIRA,[2] LUIS HUMBERTO REZENDE BARBOSA,[1]
JÚLIO CÉSAR DIAS LOPES,[3] and CLAUDIONOR NUNES COELHO JUNIOR[1]

ABSTRACT

The Tropical Biominer Project is a recent initiative from the Federal University of Minas Gerais (UFMG) and the Oswaldo Cruz foundation, with the participation of the Biominas Foundation (Belo Horizonte, Minas Gerais, Brazil) and the start-up Homologix. The main objective of the project is to build a new resource for the chemogenomics research, on chemical compounds, with a strong emphasis on natural molecules. Adopted technologies include the search of information from structured, semi-structured, and non-structured documents (the last two from the web) and datamining tools in order to gather information from different sources. The database is the support for developing applications to find new potential treatments for parasitic infections by using virtual screening tools. We present here the midpoint of the project: the conception and implementation of the Tropical Biominer Database. This is a Federated Database designed to store data from different resources. Connected to the database, a web crawler is able to gather information from distinct, patented web sites and store them after automatic classification using datamining tools. Finally, we demonstrate the interest of the approach, by formulating new hypotheses on specific targets of a natural compound, violacein, using inferences from a Virtual Screening procedure.

INTRODUCTION

CHEMOGENOMICS "knowledge space" analysis is considered, in parallel with systems biology, one of the most promising avenues for knowledge-based drug discovery. While search algorithms for ligand–target interactions have become reasonably robust, it is recognized that database integration and datamining algorithm benchmarking (Schuffenhauer and Jocaby, 2004) are actually limiting the expansion of the chemogenomics. Proposed solutions for both these problems rely on modular implementation of both data

[1] Department of Computer Science, Federal University of Minas Gerais (UFMG), Pampulha, Belo Horizonte, MG, Brazil.
[2] Fundação Oswaldo Cruz, Centro de Pesquisas René Rachou, Barro Preto, Belo Horizonte, MG, Brazil.
[3] Department of Chemistry, UFMG, Belo Horizonte, MG, Brazil.
[4] Homologix, Belo Horizonte, MG, Brazil.

access and relevant algorithms for mining these data. A third aspect to consider is the implementation of efficient management system, which supports the design of pipeline for automatic analyses. Here, we present elements of the Tropical Biominer Project that propose some answers to these problems and illustrate how to implement them in a database system dedicated to natural compounds.

The different elements of the application are built upon a Federated Database concept (Hammer et al., 1979; Heimbigner et al., 1985), and it is composed of structured, semi-structured and non-structured data sources to feed the database (information gathering using web mining tools and databases federation), and query tools that comprise simple modules such as entry visualization and editing, or more sophisticated tools like virtual screening (Fig. 1). Using the system, we succeeded in demonstrating how electronic information collected over the Internet can be aggregated to bring new insights about a specific ligand, violacein, with new hypotheses on potential targets for the molecule.

FEDERATED BIO-DATABASE SYSTEM: THE TROPICAL BIOMINER DATABASE

Today, information about plant extracts, molecules, proteins, and other biological objects is disseminated in various heterogeneous databases, and as predicted by Discala et al. (2000), the situation is still unlikely to change given the deluge of data from genomic and proteomic projects. Chemical databases, specifically, are crucial databases used in drug discovery, as data resources for virtual screening methodologies, ligand based or target based like molecular docking (Waszkowycz et al., 2001). Information maintained by the databases can cover the chemical structure of compounds and a set of attributes for the compound, which depends on the scope of the database. Most of the ligand databases—commercial like the World Drug Index (⟨www.derwent.com/products/lr/wdi/⟩), the CAS/Scifinder database (⟨www.cas.org/SCIFINDER/⟩), or the ACX database (⟨www.camsoft.com⟩), or freely available like Chembank (Schreiber, 2004) or the Open NCI database (Milne et al., 1994)—are in structured formats, although adopting different standards. More complicated is the existence of "non-structured" or "semi-structured" information on the Internet. One elegant and powerful solution to attempt to integrate such heterogeneous data is the database federation (Federated Databases [FDBs]).

Basically, the great advantage of the FDBs is to integrate, through a software layer called "federation," a set of autonomous and heterogeneous resources that can be used uniformly (Fig. 2). These resources can be distributed on different computers, each operating on different systems or with different structures. While

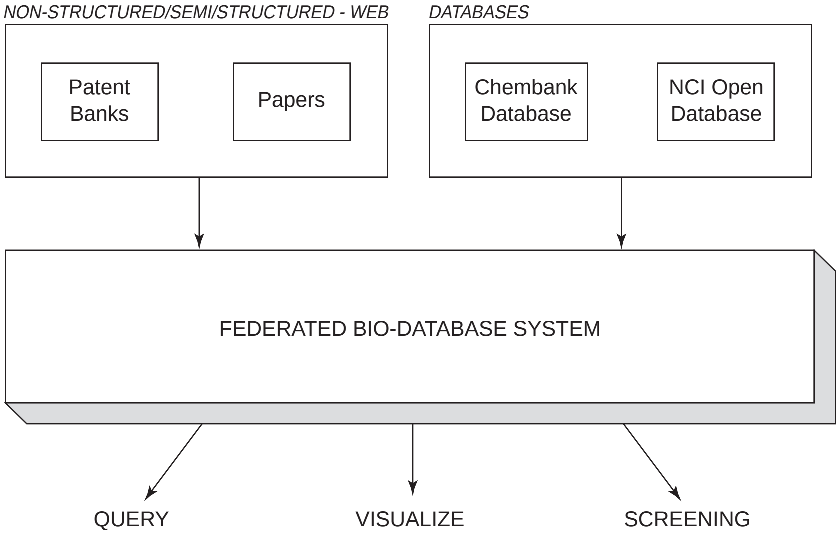


FIG. 1. Overview of the Tropical Biominer project.

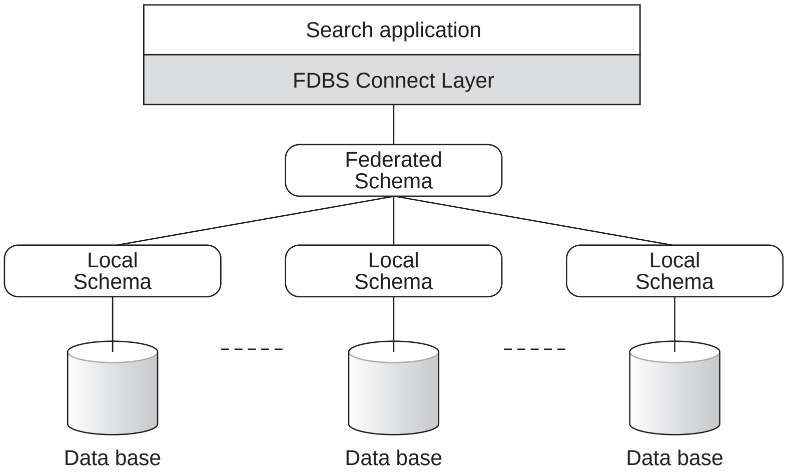


FIG. 2. Federated database model. The FDBS Connect Layer is the interface link between the applications and the Federated Databases. This layer proposes a unified Application Protocol Interface (API) to query the different databases independently of their own representations. The Federated Schema is used by any application to access the federated databases, through the FDBS Connect Layer. Commands and data are checked for syntax correctness, integrity violation rules, and access control. The entity splits and dispatches the input query into specific queries specialized for the respective databases, and integrates the results sent by the different federated databases. The Local Schema is responsible for checking the query dispatched to a local database. The query is converted to the local language, and its result is converted back to the federation schema. This entity provides the independence between the federation schema and the local schema.

it doesn't modify the "raison d'être" of the integrated resources, each keeping its own niche both in term of chemical structure exclusivity and of associated functional annotations, it allows common access to the data, solving one of the main problems of chemical databases integration, the structure-to-structure–based integration. Voigt et al. (2001) noticed this problem while integrating eight chemical databases, using unique hash code indexation. For example, one of the cited problems was that the lack of charge assignments for some compounds in the Sigma-Aldrich Catalog (⟨www.sigma-aldrich.com⟩) caused the calculation of wrong hash codes.

To initiate the project, we conceived a FDB model using a bottom-up methodology using two large chemical databases, the NCI Open Database (Monks et al., 1997) and the Chembank database maintained at Harvard University (⟨http://chembank.med.harvard.edu/⟩). The application proposes interfaces to retrieve, update, insert, and delete molecule data from distinct distributed databases. While we exemplify our model using two databases, the system is built to allow any number of databases to be connected. Using the application, the user may select which queries to perform in parallel on each desired database. Some common database table management commands, such as "drop table" or "insert table" commands, were implemented. The architecture conception was driven by several requirements, some of which are presented below:

- The access to the different database resources should be transparent for the final user. Queries and Results are filtered by the application, hence freeing the user of knowing the internal structure and representation of the underlying data.
- The FDB is able to scale up quantitatively. The schema allows the distribution the federated databases on different computers, which simplify management tasks and it provides a simple way to add new data by plugging new data servers. Furthermore, the Connect Layer can act as a task manager, dispatching jobs, which confers to the application the ability to perform distributed computation over the data.
- The query interface and the Connect Layer are Java applications running on Windows, Unix or Linux platforms.

The following components were designed and implemented:

Local base

The first task was to import the chosen databases into a SGBD tool, as both databases are provided in SDF (Structure Data File) format (Dalby et al., 1992). While these files could be used in our application, it would not provide acceptable performance rate for our application. For this, we use MS-Access, but any other SGBD could be used. Two independent databases were created, one with the NCI data and other with the ChemBank data. Data was imported using the JChemManager tool (CSIZMADIA, 2000). To manipulate the databases, we have developed a module called SearchDB, which helps in the query processing. This interface uses libraries from the Jchem modules, which supports different specialized queries for biological applications.

Local manager

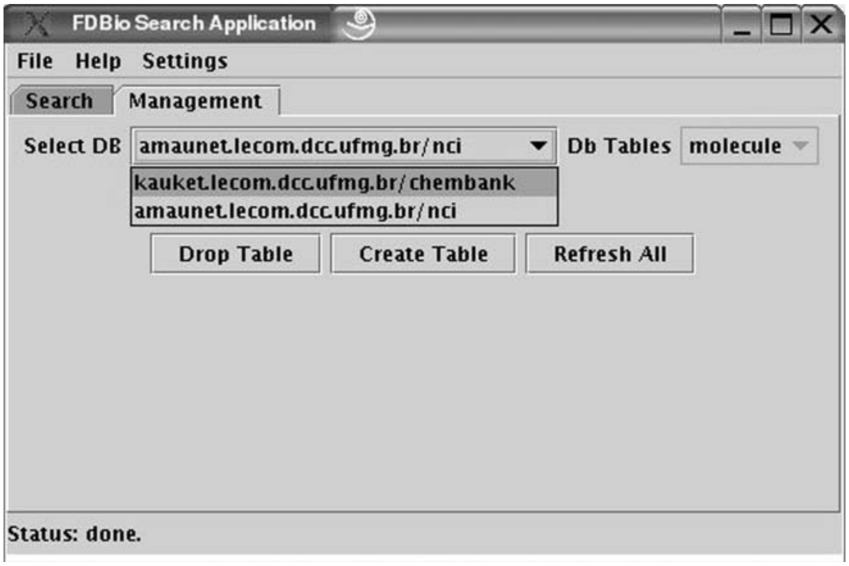
This entity receives the user query addressed to the federation, performs some checking tests and returns the compiled results. In the application, the query is stored as an XML document that contains the query parameters. From this document, local SQL queries are then generated to screen the individual databases.

FDB manager

The manager is the core of the Federated Schema. It is associated with three sub-modules that provide the distribution and integration of the federated data. These sub-modules are as follows:

1. The Query Descriptor module describes the rules necessary for dispatching the queries. It also contains the model for the query XML description.
2. The Query Store stores the provided XML queries for the federation. This entity is also responsible to merge the responses sent by the local databases and deliver the global result to the application.
3. The Local Connector provides the link between the federation and the local databases. In the proposed schema, this entity is the layer that allows the abstraction of local databases, including the technology use to access them.

Figures 3–5 illustrate different interface windows used to query and consult the data using the model.



Select DB: amaunet.lecom.dcc.ufmg.br/nci

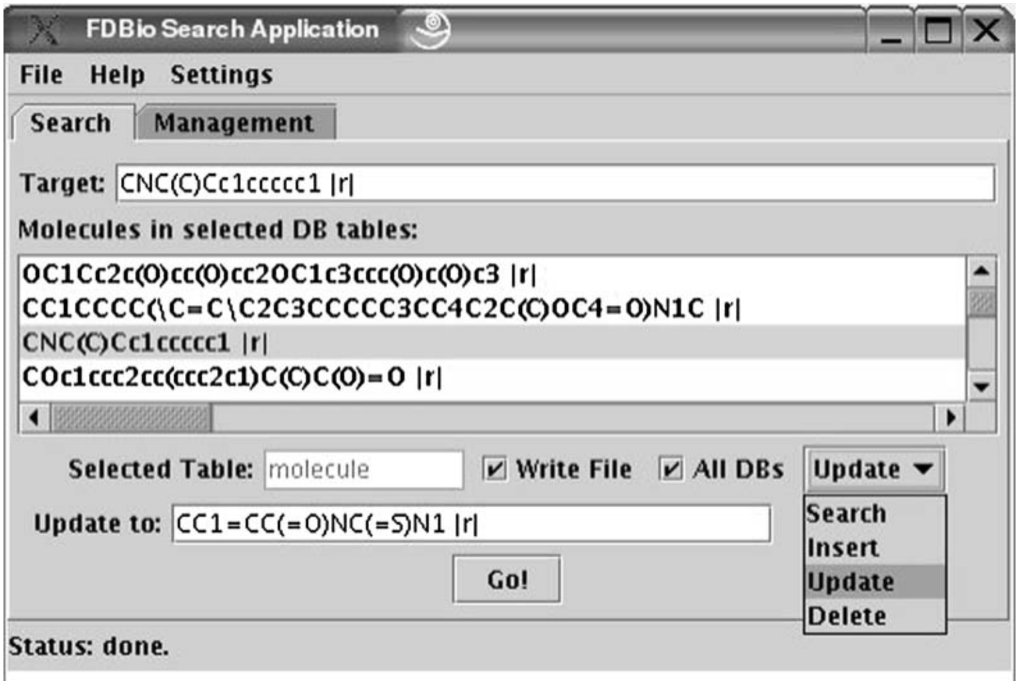
kauket.lecom.dcc.ufmg.br/chembank

amaunet.lecom.dcc.ufmg.br/nci

Db Tables: molecule

Status: done.

FIG. 3. The Management Interface is responsible for creating, destroying, and retrieving information from the available databases. It also provides graphical information about available molecule data, making it easier to perform queries as well as to manipulate a huge amount of data or very heterogeneous and independent databases (or a combination of both).



Target: CNC(C)Cc1ccccc1 |r|

Molecules in selected DB tables:

```
OC1Cc2c(O)cc(O)cc2OC1c3ccc(O)c(O)c3 |r|
CC1CCCC(\C=C\C2C3CCCCC3CC4C2C(C)OC4=O)N1C |r|
CNC(C)Cc1ccccc1 |r|
COc1ccc2cc(ccc2c1)C(C)C(O)=O |r|
```

Selected Table: molecule ☑ Write File ☑ All DBs Update

Update to: CC1=CC(=O)NC(=S)N1 |r|

Status: done.

FIG. 4. The Query Interface allows handling molecules from any database's tables and proposes statement commands (search, insert, update, delete) on the desired database.

WEB MINING AS A RESOURCE TO POPULATE THE FEDERATED BIO-DATABASE SYSTEM

Many important biotechnological advances on new drugs are published results available over the Internet or in patent banks. For example, the results of the genome project are expected to be linked to many new drug developments, as molecular and atomic researches related to genomes will contribute solutions

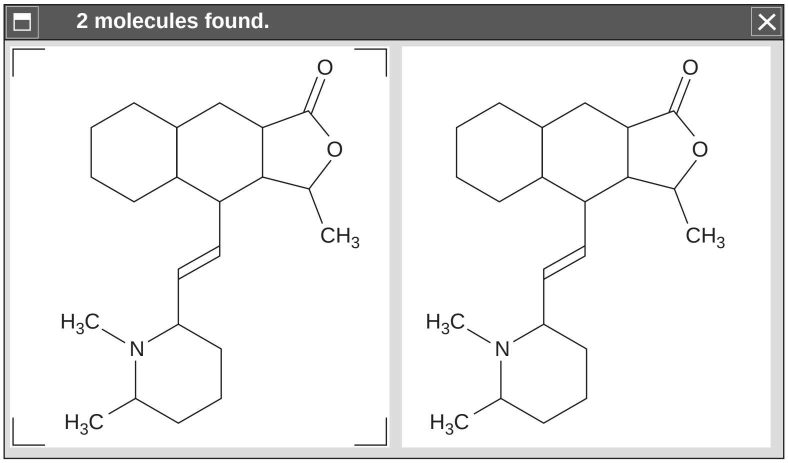

2 molecules found.

FIG. 5. Result Interface. The interface displays the structures identified on all the federated databases. The figure illustrates a molecule found in both the NCI and ChemBank Database. The interface implements the Marvin JavaBeans, which allows editing the molecular structure. A synchronization process permits the storing of the manual modification of structures into the original entry.

for many problems in areas such as medicine, agriculture, and chemistry. This research is generating a huge amount of information about these new developments, and today we observe a huge increase in research about such subjects distributed over the web.

We are developing a connection to the FBDS described earlier, based on semi-structured or non-structured sources of data (such as data retrieved from the web), using state-of-the-art web mining techniques. In the first phase, we have linked important sources of information about biotechnological advances from patent offices, since they gather information about inventions that can guide future works in biotechnology, and because patents provide a source of data in a semi-structured way. Studying these inventions allows us to look for important hints about scientific and technological advance trends. However, new information retrieval systems need to be developed to enable us to analyze this information, especially when the amount of information is so large. By connecting these systems to a local federated database, we can avoid traversing the web and deciding on where is the important information every time we have to make a query or to search for new clues, saving resources such as bandwidth and processing time. Thus, the main motivation for implementing this system is to concentrate information in a knowledge database that enables us to get important findings from previous researches in an easier way.

We have implemented a system to gather and manage information on biotechnological inventions in a local database. Our system is a web crawler that visits patent pages on the Internet and stores them in a database. Our patent crawler submits a set of queries to the United States Patent and Trademark Office (⟨www.uspto.gov/⟩) and to the European Patent Office (⟨ep.espacenet.com/⟩) sites and it stores the results in the FDB for further analysis. Our main contribution is a repository of biotechnological patents that can be processed to promote new scientific and technological advances. This information can be associated with information from other databases and it should speed up biotechnology developments.

During search in patent banks, we may retrieve several patents related to a specific subject, but not concerned with biotechnology advances. We use data mining techniques to manage this data and to classify the useful information for our purposes.

APPLICATION OF THE TROPICAL BIOMINER DATABASE: TARGET PREDICTION AND BIOLOGICAL ACTIVITY INFERENCE OF VIOLACEIN

To illustrate potential project applications, we performed some analyses on a natural compound, violacein (Bromberg and Duran, 2001; Antonio and Creczynski-Pasa, 2004). This molecule (Fig. 6, structure 4) is a violet pigment produced by *Chromobacterium violaceum*, which is widely found in the water and soil of tropical and subtropical areas. It has been previously reported that violacein has anti-tumor (Melo et al., 2000), anti-*Leshmanial* (Leon et al., 2001), antibiotic, and anti–*Trypanosoma cruzi* activities (Duran and Menck, 2001), but proposed targets do not exist. Applying simple Virtual Screening (Waszkowycz et al., 2001) methodologies to the Tropical Biominer Database, we succeeded in aggregating more information about this molecule, and we could propose new development strategies for this drug candidate.

In a first approach, we used PASS (⟨www.ibmh.msk.su/PASS/index.html⟩) method (Filimonov and Poroikov, 1996) to predict the biological spectrum activity of violacein. This analysis predicted 22 activities with a high level of confidence on a total of 34 predicted activities. From these 22 activities, two were already cited: anti-leishmanial activity (Pa = 0.513) and antiviral activity (Pa = 0.483). Among the other predicted biological activity spectrum, there are several actions that might become the basis for new applications or optimization of the substance. For example, the following activities can be cited: 5-hydroxytryptamine release stimulant (Pa = 0.792), protein kinase inhibitor (Pa = 0.683), and tumour necrosis factor–alpha release inhibitor (Pa = 0.673).

Secondly, we queried the Tropical Biominer Database (August 2004 version) in order to identify structural registered related compounds. The approach we have chosen is based on the hypothesis that compounds with the same activity share *common patterns* in their corresponding descriptors. These common patterns represented by a *hypothesis*, can be regarded as a model active structure. The target library is then scanned for structures matching this hypothesis in contrast to individual query structures. Molecular de-

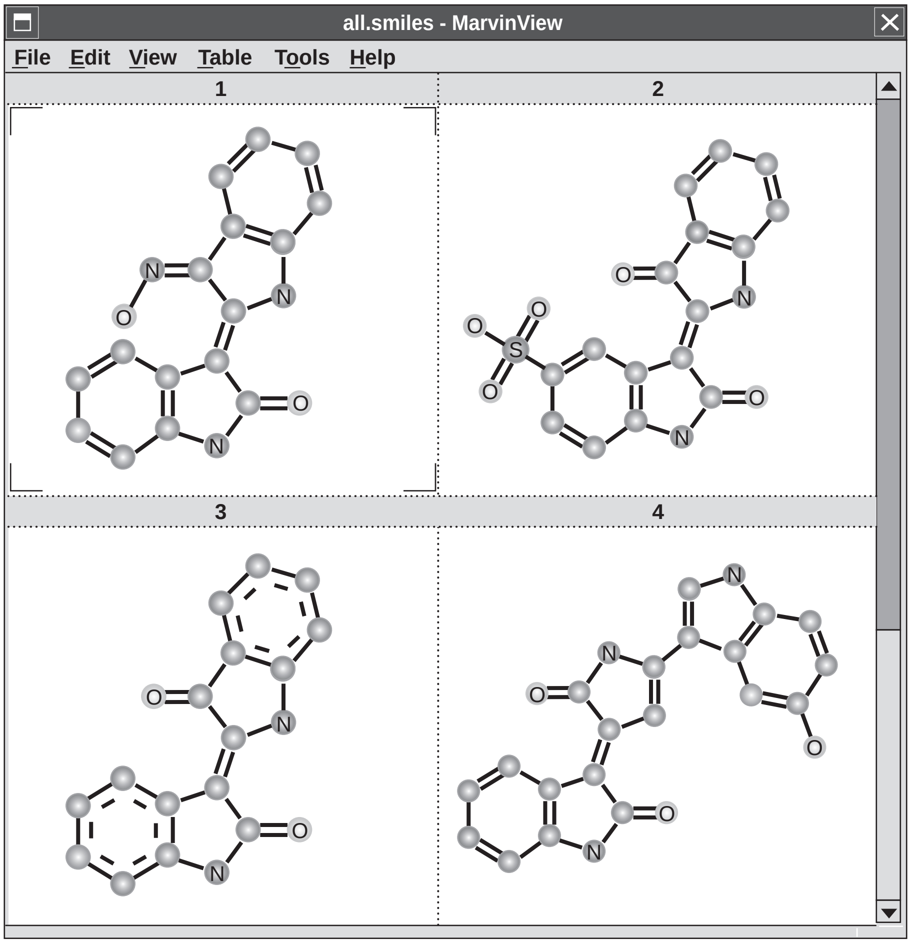


FIG. 6. Diagrams showing the chemical structures of indirubin-3-monoxime (**1**), indirubin-5′-sulphonate (**2**), indirubin (**3**), and violacein (**4**).

scriptors offer a simple feasible way to create such models and allow very fast structural queries to be applied.

Such a method was used to identify structural homologues in the Tropical Biominer database (about 300,000 molecules from Open NCI and ChemBank databases) using chemical hashed fingerprint comparison. The chemical fingerprints were generated using the ChemAxon chemo informatics Jchem package. The violacein fingerprint pattern was compared to each compounds of the database by calculating the *dissimilarity score* between them. The screening procedure was set up to accept a dissimilarity score below a threshold score defined for two dissimilarity metrics (Tanimoto, 0.4, and Euclidian, 10, metrics). Results of the database scanning lead to the selection of 54 distinct similar compounds.

By exploring functional information registered for these molecules, we focused our attention to one compound for which many data were available and registered: the indirubin (Medical Subject Headings ID C027185). This molecule is the active ingredient of the Danggui Longhui Wan, a mixture of plants that is used in traditional Chinese medicine to treat chronic diseases. It has been demonstrated that indirubin analogues and derivatives (Fig. 5) are potent inhibitors of cyclin-dependent kinases (CDKs) or the regulatory serine/threonine kinase GSK-3β (Bertrand et al., 2003; Eisenbrand et al., 2004). Based on this knowledge and considering the similar pharmaceutical profile of indirubin and violacein (PASS predictions and proved experimental activities), we decided to start complementary studies on some derivatives of the violacein, by molecular docking against CDKs structures.

DISCUSSION

We are finalizing the building of an application to visualize and to graphically search for drug, molecules, or diseases. In addition, we are finalizing the integration of the patent crawler to our federated database. To further populate the Tropical Biominer Database, we plan on adding some commercial databases to the system, and data from tropical molecules and diseases published in the scientific publications, by searching non-structured documents over the web.

ACKNOWLEDGMENTS

We acknowledge the Brazilian Research Council (CNPq), under project Tropical BioMiner, for support of this research.

REFERENCES

AMIT, P., and JAMES, A.L. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput Surv **22,** 183–236.

ANTONIO, R.V., and CRECZYNSKI-PASA, T.B. (2004). Genetic analysis of violacein biosynthesis by *Chromobacterium violaceum*. Genet Mol Res **31,** 85–91.

BERTRAND, J.A., THIEFFINE, S., VULPETTI, A., et al. (2003). Structural characterization of the GSK-3β active site using selective and non-selective ATP-mimetic inhibitors. J Mol Biol **333,** 393–407.

BROMBERG, N., and DURAN, N. (2001). Violacein biotransformation by basidiomycetes and bacteria. Lett Appl Microbiol **33,** 316–319.

CSIZMADIA, F. (2000). JChem: Java applets and modules supporting chemical database handling from web browsers. J Chem Inf Comput Sci **40,** 323–324.

DALBY, A., NOURSE, J.G., HOUNSHELL, W.D., et al. (1992). Description of several chemical-structure file formats used by computer-programs developed at Molecular Design Limited. J Chem Inf Comput Sci **32,** 244–255.

DISCALA, C., BENIGNI, X., BARILLOT, B., et al. (2000). DBcat: a catalog of 500 biological databases. Nucleic Acids Res **28,** 8–9.

DURAN, N., and MENCK, C.F.M. (2001). *Chromobacterium violaceum*: a review of pharmacological and industrial perspective. Crit Rev Microbiol **27,** 201–222.

EISENBRAND, G., HIPPE, F., JAKOBS, S., et al. (2004). Molecular mechanisms of indirubin and its derivatives: novel anticancer molecules with their origin in traditional Chinese phytomedicine. J Cancer Res Clin Oncol **130,** 627–635.

FILIMONOV, D.A., and POROIKOV, V.V. (1996). PASS: computerized prediction of biological activity spectra for chemical substances. In *Bioactive Compound Design: Possibilities for Industrial Use* (BIOS Scientific Publishers, Oxford), pp. 47–56.

HAMMER, M., and MCLEOD, D. (1979). On database management system architecture. Technical Response MIT/LCS/TM-141. Massachusetts Institute of Technology, Cambridge.

HEIMBIGNER, D., and MCLEOD, D. (1985). A federated architecture for information management. ACM Trans Off Syst **3,** 253–278.

LEON, L.L., MIRANDA, C.C., DE SOUZA, A.O., et al. (2001). Anti-leishmanial activity of the violacein extracted from *Chromobacterium violaceum*. Antimicrob Agents Chemother **48,** 449–450.

MELO, P.S., MARIA, S.S., VIDAL, B.C., et al. (2000). Violacein cytotoxicity and induction of apoptosis in V79 cells. In Vitro Cell Dev Biol Anim **36,** 639–543.

MILNE, G.W.A., NICKLAUS, M.C., DRISCOLL, J.S., et al. (1994). National Cancer Institute drug information system 3D database. J Chem Inf Comput Sci **34,** 1219–1224.

MONKS, A.P., SCUDIERO, D.A., JOHNSON, G.S., et al. (1997). The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets. Anticancer Drug Des **12,** 533–541.

SCHREIBER, S. (2004). Stuart Schreiber: biology from a chemist’s perspective. Interview by Joanna Owens. Drug Discov Today **9,** 299–303.

SCHUFFENHAUER, A., and JACOBY, E. (2004). Annotating and mining the ligand-target chemogenomics knowledge space. Drug Discov Today **2,** 190–200.

VOIGT, J.H., BIENFAIT, B., WANG, S., et al. (2001). Comparison of the NCI Open Database with seven large chemical structural databases. J Chem Inf Comput Sci **41,** 702–712.

WASZKOWYCZ, B., PERKINS, D.J., SYKES, R.A., et al. (2001). Large-scale virtual screening for discovering leads in the post-genomic era. IBM Syst J **40,** 360–376.

Address reprint requests to:
Dr. François Artiguenave
Homologix
Rua Desembargador Jorge Fontana 427/412
Belvedere—CEP 30320-670
Belo Horizonte, MG, Brazil

E-mail: artiguenave@magic.fr